**Assignment18.1**

**Problem Statement**

1. **What is the distribution of the total number of air-travelers per year**

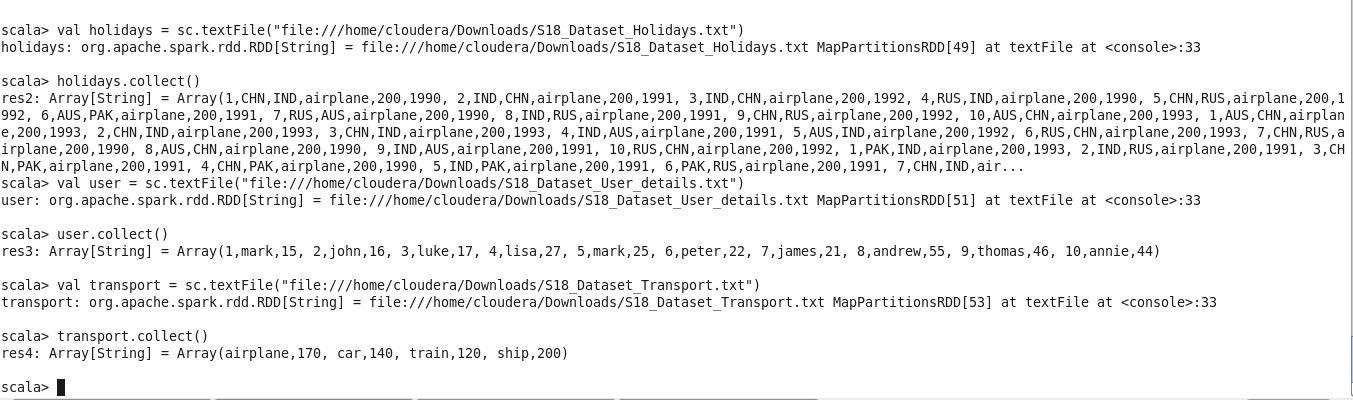
For this we have to use the datasets which contains fields

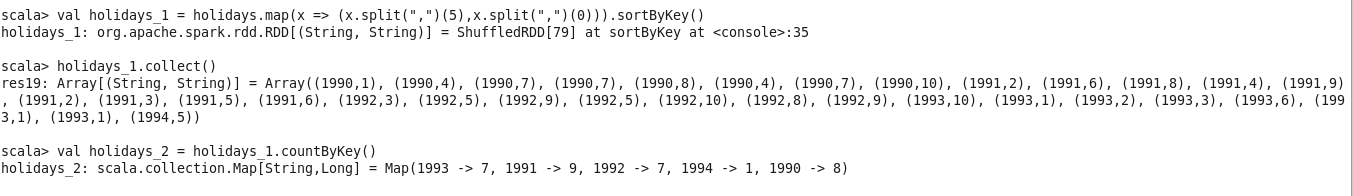
Holidays: id, source, destination, transport\_mode, distance, year

User: id, name, age

Transport: transport\_mode, cost\_per\_unit

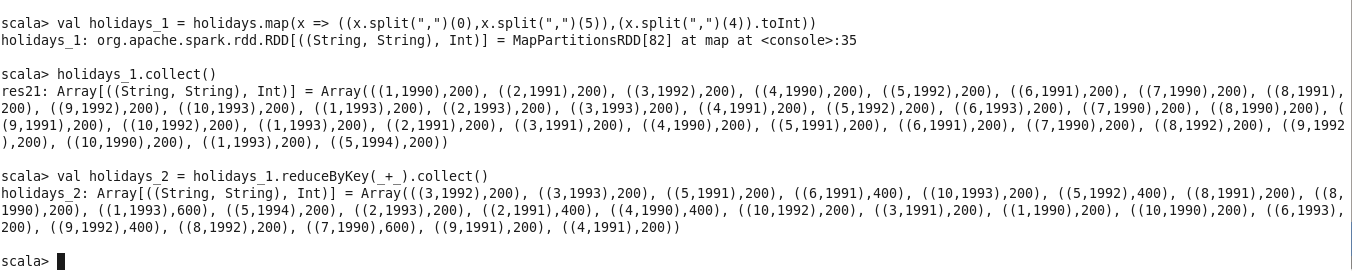
We will load all the datasets in the form of RDD





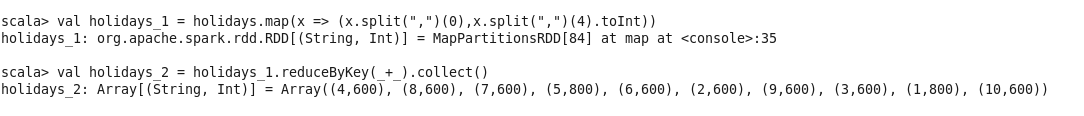
Hence by using countByKey function on key value pair we get the result as a map.

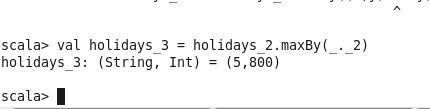
**2)** **What is the total air distance covered by each user per year**



We mapped the id and year as a key and distance as value and then used reduceByKey to get the total distance.

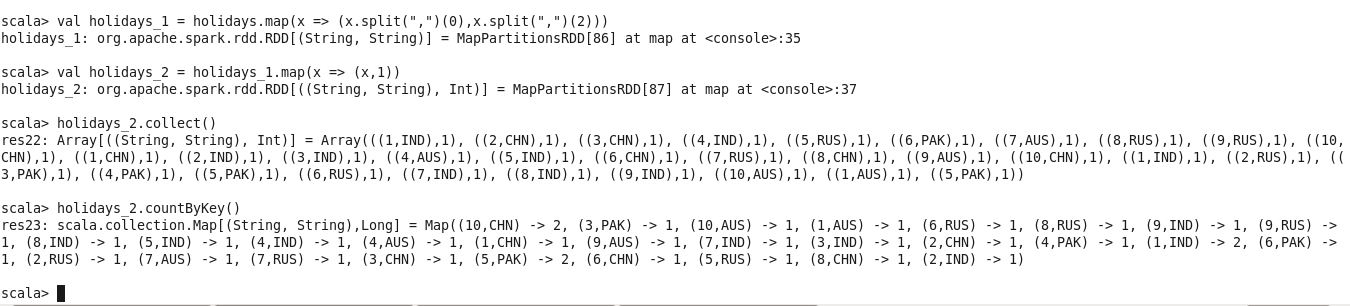
**3)Which user has travelled the largest distance till date**





We used the maxBy function to get the max of key value pair

**4) What is the most preferred destination for all users**.



We have grouped the user id and destination as a key and on that performed a count to get count of preferred destinations for all users.